APPLICATION
FOR
UNITED STATES LETTERS PATENT

Applicant:

Jochen C. Beintner, et al.

For:

"Selfaligned S/D FinFet Process Flow"

Docket: **YOR920030380US1**

INTERNATIONAL BUSINESS MACHINES CORPORATION
ARMONK, NEW YORK 10504

I HEREBY CERTIFY THAT THIS CORRESPONDENCE IS BEING DEPOSITED WITH THE UNITED STATES POSTAL SERVICE AS EXPRESS MAIL IN AN ENVELOPE ADDRESSED TO: U.S. PATENT AND TRADEMARK OFFICE, WASHINGTON, D.C. 20231 THE APPLICANT AND/OR ATTORNEY REQUESTS THE DATE OF DEPOSIT AS THE FILING DATE.

Express Mail No: ER611131618US
Date of Deposit: 12/09/03
Name of Person Making Deposit: Eric W. Petraske
Signature: E W Petraske

Selfaligned Source/Drain FinFET Process Flow

TECHNICAL FIELD

The field of the invention is that of fabricating field effect transistors having a body extending perpendicular to the semiconductor substrate between horizontally disposed source and drain regions, referred to as a "FinFET".

BACKGROUND OF THE INVENTION

Metal-Oxide-Semiconductor field effect transistor (MOSFET) technology is the dominant electronic device technology in use today. Performance enhancement between generations of devices is generally achieved by reducing the size of the device, resulting in an enhancement in device speed. This is generally referred to as device "scaling".

Ultra-large-scale integrated (ULSI) circuits generally include a multitude of transistors, such as more than one million transistors and even several million transistors that cooperate to perform various functions for an electronic component. The transistors are generally complementary metal oxide semiconductor field effect transistors (CMOSFETs) which include a gate conductor disposed between a source region and a drain region. The gate conductor is provided over a thin gate oxide material. Generally, the gate conductor can be a metal, a polysilicon, or polysilicon/germanium

(Si.sub.x Ge.sub.(1-x)) material that controls charge carriers in a channel region between the drain and the source to turn the transistor on and off. The transistors can be N-channel MOSFETs or P-channel MOSFETs.

In bulk semiconductor-type devices, transistors such as MOSFETs, are built on the top surface of a bulk substrate. The substrate is doped to form source and drain regions, and a conductive layer is provided between the source and drain regions. The conductive layer operates as a gate for the transistor; the gate controls current in a channel between the source and the drain regions. As transistors become smaller, the body thickness of the transistor (or thickness of depletion layer below the inversion channel) must be scaled down to achieve superior short-channel performance.

As MOSFETs are scaled to channel lengths below 100 nm, conventional MOSFETs suffer from several problems. In particular, interactions between the source and drain of the MOSFET degrade the ability of the gate to control whether the device is on or off. This phenomenon is called the "short-channel effect".

Silicon-on-insulator (SOI) MOSFETs are formed with an insulator (usually, but not limited to, silicon dioxide) below the device active region, unlike conventional "bulk" MOSFETs, which are formed directly on silicon substrates, and hence have silicon below the active region.

Conventional SOI-type devices include an insulative substrate attached to a thin-film semiconductor substrate that contains transistors similar to the

MOSFETs described with respect to bulk semiconductor-type devices. The insulative substrate generally includes a buried insulative layer above a lower semiconductor base layer. The transistors on the insulative substrate have superior performance characteristics due to the thin-film nature of the semiconductor substrate and the insulative properties of the buried insulative layer. In a fully depleted (FD) MOSFET, the body thickness is so small that the depletion region has a limited vertical extension, thereby eliminating link effect and lowering hot carrier degradation. The superior performance of SOI devices is manifested in superior short-channel performance (i.e., resistance to process variations in small size transistors), near-ideal subthreshold voltage swing (i.e., good for low off-state current leakage), and high saturation current. SOI is advantageous since it reduces unwanted coupling between the source and the drain of the MOSFET through the region below the channel. This is often achieved by ensuring that all the silicon in the MOSFET channel region can be either inverted or depleted by the gate (called a fully depleted SOI MOSFET). As device size is scaled, however, this becomes increasingly difficult, since the distance between the source and drain is reduced, and hence, they increasingly interact with the channel, reducing gate control and increasing short channel effects (SCE).

The double-gate MOSFET structure is promising since it places a second gate in the device, such that there is a gate on either side of the channel. This allows gate control of the channel from both sides, reducing SCE. Additionally, when the device is turned on using both gates, two conduction ("inversion") layers are formed, allowing for more current flow. An

extension of the double-gate concept is the "surround-gate" or "wraparound-gate" concept, where the gate is placed such that it completely or almost-completely surrounds the channel, providing better gate control.

In a double gate field effect transistor (FinFET), the device channel comprises a thin silicon fin standing on an insulative layer (e.g. silicon oxide) with the gate in contact with the sides of the fin. Thus inversion layers are formed on the sides of the channel with the channel film being sufficiently thin such that the two gates control the entire channel film and limit modulation of channel conductivity by the source and drain.

The double gates on the channel fin effectively suppress SCE and enhance drive current. Further, since the fin is thin, doping of the fin is not required to suppress SCE and undoped silicon can be used as the device channel, thereby reducing mobility degradation due to impurity scattering. Further, the threshold voltage of the device may be controlled by adjusting the work function of the gate by using a silicon-germanium alloy or a refractory metal or its compound such as titanium nitride.

Generally, it is desirable to manufacture smaller transistors to increase the component density on an integrated circuit. It is also desirable to reduce the size of integrated circuit structures, such as vias, conductive lines, capacitors, resistors, isolation structures, contacts, interconnects, etc. For example, manufacturing a transistor having a reduced gate length (a reduced width of the gate conductor) can have significant benefits. Gate conductors with reduced widths can be formed more closely together, thereby

increasing the transistor density on the IC. Further, gate conductors with reduced widths allow smaller transistors to be designed, thereby increasing speed and reducing power requirements for the transistors.

Heretofore, lithographic tools are utilized to form transistors and other structures on the integrated circuit. For example, lithographic tools can be utilized to define gate conductors, active lines conductive lines, vias, doped regions, and other structures associated with an integrated circuit. Most conventional lithographic fabrication processes have only been able to define structures or regions having a dimension of 100 nm or greater.

In one type of conventional lithographic fabrication process, a photoresist mask is coated over a substrate or a layer above the substrate. The photoresist mask is lithographically patterned by providing electromagnetic radiation, such as ultraviolet light, through an overlay mask. The portions of the photoresist mask exposed to the electromagnetic radiation react (e.g. are cured). The uncured portions of the photoresist mask are removed, thereby transposing the pattern associated with the overlay to the photoresist mask. The patterned photoresist mask is utilized to etch other mask layers or structures. The etched mask layer and structures, in turn, can be used to define doping regions, other structures, vias, lines, etc.

As the dimensions of structures or features on the integrated circuit reach levels below 100 nm or 50 nm, lithographic techniques are unable to precisely and accurately define the feature. For example, as described above, reduction of the width of the gate conductor (the gate length)

associated with a transistor or of the active lines associated with an SOI transistor has significant beneficial effects. Future designs of transistors may require that the active lines have a width of less than 50 nanometers.

Double gate SOI MOSFETs have received significant attention because of its advantages related to high drive current and high immunity to short channel effects. The double-gate MOSFET as able to increase the drive current because the gate surrounds the active region by more than one layer (e.g., the effective gate total width is increased due to the double gate structure). However, patterning narrow, dense active regions is challenging. As discussed above with respect to gate conductors, conventional lithographic tools are unable to accurately and precisely define active regions as structures or features with dimensions below 100 nm or 50 nm.

Thus, there is a need for an integrated circuit or electronic device that includes smaller, more densely disposed active regions or active lines. Further still, there is a need for a ULSI circuit which does not utilize conventional lithographic techniques to define active regions or active lines. Even further still, there is a need for a non-lithographic approach for defining active regions or active lines having at least one topographic dimension less than 100 nanometers and less than 50 nanometers (e.g., 20-50 nm). Yet further still, there is a need for an SOI integrated circuit with transistors having multiple sided gate conductors associated with active lines having a width of about 20 to 50 nm.

SUMMARY OF THE INVENTION

The present invention is directed to a process for fabricating FinFET transistor structures, in which the source and drain are selfaligned to the gate.

A feature of the invention is that the gate is formed before source and drain contacts are made and the gate is encapsulated by a dielectric material.

Another feature of the invention is that the selfalignment provides better control of fringe capacitance and external resistance than prior art methods.

Another feature of the invention is that the gate is deposited in an aperture formed in a blanket dielectric.

Yet another feature of the invention is that the gate is fully silicided.

BRIEF DESCRIPTION OF THE DRAWINGS

Figures 1 - 11 illustrates steps in the inventive process.

Figures 12 - 16 illustrate steps in an optional gate silicidation process.

DETAILED DESCRIPTION

Figures 1A and 1B show cross sections of a set of fins 30 that will become the fins of a FinFET. As used herein, the term "set" means one or more; i.e. a FinFET may have one or more fins. Figure 1C shows the location of the cross sections of Figures 1A and 1B. In this example, the four fins shown will be controlled by a common gate. Those skilled in the art will be aware that separated gates could be formed to control one or more fins, if desired. The Figure shows the result of conventional preliminary steps, well known to those skilled in the art, of forming the silicon fins for a FinFET.

Narrow fin structures in silicon or silicon on insulator (SOI) can be fabricated in different ways, e.g. by optical lithography followed by different trimming techniques (resist trimming, hard mask trimming, oxidation trimming). These processes are based on width reduction of the mask by plasma etch or wet etch, or by material consumption of the fin by oxidation)), by E-beam lithography or by sidewall image transfer processes.

In this example illustrated, the sidewall image transfer process is used as the method to structure narrow fins in SOI. Figure 1 shows a bulk wafer 10, having a buried oxide (BOX) 20 with an SOI layer 30 of 70nm (Possible range of the SOI is ~10nm to 200nm, but not limited to that range). The surface of layer 30 has been oxidized to form 300Å of thermal oxide 32 (Preferred range 50Å-500Å). Alternatively, an oxide can also be deposited using any kind of CVD processes.

The following discussion illustrates a conventional method, well known to

those skilled in the art, of fabricating the structure shown in Figure 1. Other methods may also be used. These initial steps are not illustrated in the Figures to avoid unnecessary detail. Initially, 1500Å (Preferred range 500Å -3000Å) of amorphous silicon were deposited on the wafer surface that will be formed into the fins (oxide layer 32 on top of fin layer 30) by CVD or sputter processes, followed by the deposition of 500Å (Preferred range 100Å -2000Å) of CVD oxide as a hardmask. Optical lithography and RIE etch processes are used to structure the oxide hardmask and the amorphous silicon layer, stopping on the oxide layer 32 on top of the SOI. Then a 200Å (Preferred range 50Å B 500Å) nitride layer (not shown) is deposited conformally using a CVD process followed by a RIE etch process to form SiN spacers on the side of the amorphous silicon.

The amorphous silicon is then removed with a plasma etch or wet etch leaving nitride spacer structures behind. The spacer structures are used as a hardmask to structure the oxide 32 underneath and can be removed afterwards by oxide and silicon selective plasma etches or wet etches (e.g. hot phosphoric acid). The structured oxide 32 is then used as a hardmask to etch the silicon fins 30 in the SOI layer. Next, a sacrificial oxide is thermally grown to remove RIE damage from the silicon fin surface and to act as a screen oxide for fin body doping implants that can be processed at this point. Fin body doping implants are not necessary to make the FinFET device work, but can be useful to set FinFET Vt.

The sacrificial oxide is removed by a wet etch, followed by a preclean and gate oxide processing using thermal oxidation or CVD deposition processes.

A specific example of the process described above is shown in copending patent application Attorney Docket Number YOR920030433US1, assigned to the assignee hereof and incorporated herein by reference and omitted from this description for simplicity.

The result of these preliminary steps is shown in Figure 1. Figure 1C shows a top view of a sample structure indicating cross sections 1A and 1B shown in Figures 1A and 1B, respectively. For convenience in explanation, the top of Figure 1C will be referred to as North, with other directions corresponding. Thus, Figure 1A is a cross section taken at the North end of the fins, looking north. The central portion of Figure 1C will be the location of the self-aligned gate constructed according to the invention.

Referring now to Figure 2, a polysilicon layer 40 of 1500Å (Preferred range 500Å to 3000Å, depending on total fin height) is deposited using a CVD process and then planarized by a CMP or planarizing coating/etchback processes to improve the process window of the gate lithography step later in the process. Optional poly pre-doping to adjust the gate workfunction for NFETs and PFETs is followed by a low temperature CVD deposition of 1200Å nitride 45 that forms a protective cap.

As explained below, the total height of a hardmask formed by layer 45 needs to be greater than silicon fin 30 height, plus the oxide 32 on the fin, plus a process margin. Before the nitride deposition, an optional oxide layer 42 (Preferred range 20Å-500Å) can be deposited to act as a stress buffer between the nitride and the polysilicon.

Nitride 45 is structured using optical lithography (the area shown as box PC in Figure 2D), e-beam lithography or sidewall image transfer processes, and RIE to form a poly structure extending E-W that will define the selfaligned gate. With this structure as a nitride hardmask, poly 40 is etched by RIE down to BOX 20 to define the gate, leaving the fins 30 standing exposed in the S/D area. The result of the litho step is that Figure 2A shows the same view as Figure 1A, while Figure 2B shows the poly structure capped by a nitride hardmask. Figure 2C shows a view looking E along the poly structure. The fin 30 extends horizontally in this view and the gate extends perpendicular to the plane of the paper, with the plane of the cross section within the poly structure and outside the fin. If a gate width that is sub-litho is desired, trimming of the nitride hardmask 45 on top of the polysilicon can be done by resist trimming techniques or by nitride wet etch or dry etch processes.

The next steps after the poly gate 40 is formed are 30Å gate sidewall oxidation (Preferred range 0Å to 100Å), 50Å CVD oxide liner deposition (Preferred range 0Å to 500Å) and implantations to process conventional halo and extension implants, well known to those skilled in the art. Then an encapsulating CVD nitride layer 50 is deposited with a thickness of 500Å (Preferred range 50Å to 1000Å) as shown in Figure 3. (The notation used is that a thickness of 0 indicates that the layer may be omitted.) The nitride has a thickness on the vertical edges of gate 40 optimized for the fringe capacitance/external resistance trade off and to offset the Source/Drain implants from the gate to adjust for source/drain dopant diffusion during

later thermal processes. Depending on the distance between each fin, the nitride may fill up the space between the fins completely, as shown in Figure 3A. In Figure 3 and the following figures, Figures n–A and n–B will have the same orientation as Figures 1A and 1B, respectively.

Since the nitride has a thickness that is ten times that of the gate sidewall oxide and oxide liner, these three layers are shown schematically by layer 50, to avoid a confusing multiplicity of lines in the Figure. Figure 3B shows layer 50 as resting on layer 45. Figure 3C shows layer 50 covering the top and sides of nitride 45 and the top and sides of the fin 30, extending N-S along fin 30 over the full height of the fin.

A CVD oxide 60 is then deposited covering the entire structure and planarized by CMP or any other planarization technique, preferably stopping on the nitride layer 50 on top of the gate nitride cap 45. The cap is the remaining nitride from the nitride hardmask on top of the polysilicon gate. The result is shown in Figure 4, with the oxide 60 covering the nitride 50 in Fig 4A and extending up to the top of the nitride cap in Fig. 4C, filling the area up to the level of the top of the nitride 50 in Figure 4C. Figure 4B is the same as Figure 3B, since oxide 60 is removed down to the top of nitride 50.

Figure 5 shows the result of a lithography process that structures the area for source/drain formation followed by an anisotropic, nitride selective oxide etch stopping on the bottom nitride 50. Figure 5D shows a rectangle 63 that covers the fins in the example, ending before it reaches the limit of the litho

rectangle that defines the gate and a rectangle 61 that represents the gate structure. Figure 5A shows the fins at the N end, in which there is an opening 62 between the remaining portion of the oxide fill. Figure 5C shows a portion of the same aperture, stopping in both Figures 5A and 5C on nitride 50. Figure 5B is unchanged, because the aperture is not as wide as the view in this cross section and the oxide has been planarized to the top of the nitride 50 in Figure 5B.

The purpose of oxide 60 is to protect etch defined source/drain areas that are isolated from each other by the oxide, such as NFETs in the following step. The oxide can be compared with an oxide deposited and structured to form contacts, as is done in the back end wiring.

In Figure 6, the nitride layer 50 that was deposited in Figure 3 is anisotropically etched selective to oxide to remove material from the top and to form spacers on the two sides of the gate 40, (Figure 6C). The etch is continued with an overetch of the nitride spacer sufficient to clean up the sidewalls of the fins (Fig. 6A), showing the silicon 30 of the fins exposed. The nitride cap 45 on top of the poly gate 40 needs to have a thickness sufficient to withstand the overetch, so that Fig. 6B shows a remaining nitride cap 45.

At this point an optional second spacer material can be deposited, e. g. CVD oxide (Preferred range 50Å - 1000Å). The oxide spacer can be formed by RIE selective to nitride and silicon and the spacer can be removed by lithographic area definition and plasma or wet etch in certain areas where it

is not required. The second spacer can also be formed before the nitride spacer is etched, resulting in the following sequence: oxide deposition on top of nitride, oxide spacer etch and area selective removal, nitride spacer etch. Bracket 52 indicates schematically the extra width of the extra oxide spacer.

If the optional oxide is chosen, the spacer will have two layers - nitride and oxide. Fig. 6C shows spacer 50 extending in the E-W direction to separate gate 40 from the S/D contacts.

The second spacer can be used to adjust for different diffusivity of Arsenic in the NFET and Boron in the PFET. As Boron diffuses faster, the source/drain offset from the gate in the PFETs needs to be larger. Accordingly, the second spacer is deposited on PFET structures to provide the required extra spacing..

If the optional oxide spacers are not used, the oxide 32 on top of the fin silicon can be removed at this point. This is an option, not a necessary step. Figures 7A and 7C show the removal of oxide 32 and also an aperture cut into the BOX as a result of the oxide etch. Protecting the BOX to avoid this aperture is not necessary, as it does not affect the operation of the FinFET. Oxide 32 remains in Figure 7B, as it is covered by the poly 40.

Figure 8 shows the beginning of the selfaligned source/drain formation. There are several different options to form the source/drain elements of the transistors:

1. Deposit undoped CVD polysilicon and planarize it stopping on the oxide. 2. Grow selective epitaxial silicon and planarize it at the same level. 3. Grow epitaxial silicon between the fins (Preferred range 50Å B 500Å, depending on fin to fin distance), process a silicide on the grown silicon, deposit a contact liner (e.g. TiN) and then fill up with metal (e.g. CVD tungsten) and planarize. Instead of silicon it is also possible to use silicon-germanium. The last option (Metal fill) is the best for device performance, since it has low external resistance. The first option (Polysilicon fill) is the cheapest. It is also possible to process the source/drain regions of NFETs and PFETs separately, giving the option of using in situ doped processes for the poly deposition and the selective epitaxy, meaning n-doped silicon for the NFET and p-doped silicon for the PFET. In this case the extensions can also be formed by solid phase outdiffusion from the silicon.

Figure 8A shows the formation of S/D material 70, filling the aperture. Fig. 8B shows a thinner layer reflecting the height of the oxide 60. Fig. 8C shows the S/D material covering the gate 40 (which is encapsulated by the nitride liner).

Figure 9 shows the result of recessing the source/drain material 70 deposited in the previous step by a plasma etch or wet etch to a height that is still covering the top of the fin 30. The purpose of this recess is to reduce the source/drain capacitance by removing the S/D material 70 where it does not contribute to a reduction of the external resistance. Ideally the recess would stop on top of the fin .

Figure 9A shows a solid block of recessed poly providing a large surface for placing a contact to the S/D. Figure 9C shows the gate separated from the S/D by spacer 50. Figure 9B shows the remaining portion of nitride 45 as the top layer.

In Figure 10, the nitride cap 45 on top of the gate is removed by an anisotropic RIE etch, exposing the top surface of the poly gate 40.

In Figure 11, the source and drain blocks, also the gate, are doped by an ion implant, separately for NFET and PFET, followed by a silicidation process (Ni, Co, Y) that form a silicide on the S/D blocks and also on the top of the gate 40. At this point the process could continue with a standard integrated circuit fabrication process, starting with CVD oxide deposition and formation of contacts and metal wiring.

The following material describes an embodiment that offers the integration of a fully silicided gate process into the selfaligned source/drain process flow.

A 100Å conformal CVD nitride film 82 (Preferred range 50Å B 500Å) is deposited after the previous silicide step, covering the FinFETs and oxide 60, Figure 12

Then a CVD oxide 92 is deposited on top of the nitride 82. The oxide film is thicker than the height difference between gate and source/drain area. The

oxide is planarized by CMP or any other planarization technique to the height of oxide 60 and nitride 82, Figure 13.

Exposed portions of nitride 82 on top of oxide 92 are stripped (e.g. hot phosphoric acid), then oxide 92 is recessed by plasma etch back or wet etch selective to nitride so that the upper part of the gate is cleared from the oxide. Another process option would be to do an oxide CMP selective to nitride, stopping on the gate nitride cap (Figure 14).

In Figure 15 the nitride cap is removed by plasma etch or by wet etch extending down along the gate sidewalls formed by nitride 50 to BOX 20, leaving a narrow aperture 41 visible in Figure 15C.

In Figure 16 the gate is fully silicided, extending vertically down to BOX 20, preferably using Ni, while the fins and S/D are protected by the CVD oxide 92. Co is an optional choice for silicidation. In this case of a full silicidation of the gate, an additional Ni deposition and anneal is necessary to convert the entire gate poly into silicide.

The process optionally then continues with a standard integrated circuit formation process and back end processes in other portions of the circuit, as described above.

Those skilled in the art will appreciate that the various deposition and etching steps are conventional. The use of silicon is not required and any semiconductor may be used. Different combinations of etch-susceptible and

etch-resistant materials may be used. Conventional planar FETs may be included in the circuit, if the designer chooses.

While the invention has been described in terms of a single preferred embodiment, those skilled in the art will recognize that the invention can be practiced in various versions within the spirit and scope of the following claims.